This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

Genome-wide tuning of protein expression levels to rapidly engineer microbial traits

Journal:	*ACS Synthetic Biology*
Manuscript ID	sb-2015-001337.R1
Manuscript Type:	Article
Date Submitted by the Author:	16-Oct-2015
Complete List of Authors:	Freed, Emily; University of Colorado Boulder, Chemical and Biological Engineering Winkler, James; University of Colorado Boulder, Chemical and Biological Engineering Weiss, Sophie; University of Colorado Boulder, Chemical and Biological Engineering Garst, Andrew; University of Colorado, Boulder, Chemical and Biological Engineering Mutalik, Vivek; Lawrence Berkeley National Laboratory, Physical Biosceinces Division Arkin, Adam; University of California, Bioengineering Knight, Rob; University of Colorado, Boulder, Dept. of Chemistry and Biochemistry Gill, Ryan; Colorado University Boulder, Department of Chemical and Biological Engineering

SCHOLARONE™
Manuscripts

Genome-wide tuning of protein expression levels to rapidly engineer microbial traits

Emily F Freed[1,*], James D Winkler[1], Sophie J Weiss[1], Andrew D Garst[1], Vivek K Mutalik[2,3], Adam P Arkin[2,3], Rob Knight[4,5,+], Ryan T Gill[1]

1. Department of Chemical and Biological Engineering, University of Colorado, Boulder, Colorado, USA.
2. Lawrence Berkeley National Laboratory, Physical Biosciences Division, Berkeley, California, USA.
3. Department of Bioengineering, University of California, Berkeley, Berkeley, California, USA.
4. Department of Chemistry and Biochemistry and BioFrontiers Institute, University of Colorado, Boulder, Colorado, USA.
5. Howard Hughes Medical Institute, Boulder, Colorado, USA.

* Current affiliation: National Bioenergy Center, National Renewable Energy Laboratory, Golden, Colorado, USA.

\+ Current affiliation: Departments of Pediatrics and Computer Science & Engineering, University of California, San Diego, La Jolla, California, USA.

Abstract

The reliable engineering of biological systems requires quantitative mapping of predictable and context-independent expression over a broad range of protein expression levels. However, current techniques for modifying expression levels are cumbersome and are not amenable to high-throughput approaches. Here we present major improvements to current techniques through the design and construction of *E. coli* genome-wide libraries using synthetic DNA cassettes that can tune expression over a ~10^4 range. The cassettes also contain molecular barcodes that are optimized for next-generation sequencing, enabling rapid and quantitative tracking of alleles that have the highest fitness advantage. We show these libraries can be used to determine which genes and expression levels confer greater fitness to *E. coli* under different growth conditions.

Keywords: Recombineering, Illumina sequencing, directed evolution, genome-wide expression library, genotype-phenotype mapping

Advances in DNA synthesis and sequencing have resulted in the development of a range of new methods and approaches for engineering biological systems at throughputs several orders of magnitude beyond prior efforts. Such approaches employ multiplex oligomer synthesis and recombineering technologies to construct precisely designed libraries containing billions of mutations on laboratory timescales.(*1, 2*) These approaches have been demonstrated in a broad range of contexts, including improving the production of various compounds(*3, 4*) and engineering tolerance to a range of industrially relevant conditions.(*3, 5*) In these cases it was necessary to initiate combinatorial library generation on a set of target genes expected to have an effect on the phenotype of interest. However, the genetic basis of many relevant phenotypes remains poorly understood.(*6*) Therefore, methods for rapidly identifying such genes have been an area of intense research effort.

Trackable multiplex recombineering (TRMR) was developed to address this need.(*2*) Specifically, TRMR provides an approach for rapidly generating genome-scale libraries of gene overexpression or downregulation that can be easily tracked using molecular barcodes.(*2*) In the original TRMR libraries, overexpression was achieved by placing a strong promoter and RBS in front of each gene in the *E coli* genome. Downregulation was achieved by replacing each native RBS with an inert sequence, resulting in a decrease in translation initiation. Each allele was tracked using molecular barcodes and microarray. Changes in barcode frequency

after cells were subjected to a selective pressure were used to determine the fitness of each allele. By using TRMR libraries, it has been possible to identify alleles that confer a fitness advantage under various medically- and industrially- relevant conditions.(*2, 7, 8*)

While such studies have proven useful, subsequent efforts to further characterize genes identified by TRMR or to engineer more complex and combinatorial phenotypes revealed several design limitations. Specifically, the synthetic parts used to build the original TRMR libraries were not standardized, and thus their use may result in inconsistent expression levels across targeted genes.(*7, 8*) Furthermore, the original TRMR libraries only included "on" or "off" expression states. While the up and down libraries were successfully used to determine changes in expression level that result in increased fitness, these large changes in expression may also be detrimental. Indeed, when engineering a synthetic pathway, the activity of the pathway is dependent on expressing many of the component proteins in a very narrow expression range.(*9*) Methods for combinatorial genome engineering therefore require synthetic cassettes that will not only span a range of expression levels but also will have predictable and context-independent effects on expression.

To address these issues, we designed a new synthetic strategy for TRMR, called tunable TRMR (T^2RMR), that combines standardized promoter and ribosome binding site (RBS) sequences that predictably modify expression within twofold of the target expression level and work over a wide range of expression levels in a context free manner.(*10*) The new libraries also contain barcodes optimized for Illumina sequencing to take advantage of the improved dynamic range of sequencing compared to microarrays.

RESULTS AND DISCUSSION

Design of T^2RMR cassettes

Tunable TRMR (T^2RMR) employs a synthetic DNA (synDNA) cassette that combines modules for i) integration into the genome via recombineering, ii) inducible control of gene expression (promoter), and iii) removing context-specific effects on expression (dual RBS) (Figure 1a). A synDNA cassette was integrated in front of every gene in the *E. coli* genome to replace the native promoter and RBS with synthetic versions. To eliminate context-dependence of gene expression across all genes/proteins, we used a bicistronic design (BCD) for ribosome binding that has been shown to give significantly more consistent expression when tested with a variety of genes.(*10*) We chose four BCDs of varying strengths to employ in our T^2RMR libraries, resulting in four different base expression levels ("off", "low", "intermediate", and "high").

To fine-tune expression levels, we additionally placed an inducible LacI-regulated promoter(*10*) in front of the BCD sequences. By varying the amount of inducer (IPTG), each of the four libraries can be tuned over a wide range. Additional sequences were included in the synDNA cassettes to avoid unintended upstream transcription or translation and downstream interference with coding sequences. These sequences are standardized and well-defined and have been validated to reduce context specific effects between genes.(*10*) Finally, each cassette also contains a selectable marker and two molecular barcodes to track both BCDs and genes (Figure 1a).

Confirming T^2RMR cassette function on a single gene scale

We tested the performance of the T^2RMR synDNA cassette design by recombineering each of the four cassettes upstream of the native *lacZ* locus in *E. coli*. Insertion of the cassettes into the correct genomic locus was verified by both PCR and sequencing. Strains containing each of the four BCDs were grown separately in differing concentrations of IPTG (0.01 mM, 0.125 mM, or 1 mM) or in no IPTG and then β-galactosidase (the protein product of the *lacZ* gene) activity was measured for each. The results showed that the "off" cassette produced almost no β-galactosidase activity, while the other three cassettes produced differing amounts of β-galactosidase activity in the expected rank order for the BCDs. Furthermore, the amount of activity could be modulated by changing the concentration of IPTG (Figure 1b). Overall, we observed an ~10^4-fold activity range.

As a second approach to confirm the constructs were functioning as predicted, whole cell lysates from each of the four strains grown in 1 mM IPTG were submitted for mass spectrometric analysis to the CU BioFrontiers Institute Mass Spectrometry and Proteomics core facility. Spectral counting(*11, 12*) was used to determine the relative abundance of proteins. Similar to the β-galactosidase assay, the "off" cassette produced almost no spectral counts, while the other three cassettes produced increasing numbers of spectral counts based on BCD strength, revealing a ~10^2 range in expression levels (Figure 1c). The β-galactosidase expression level results combined with previous validation that BCDs result in much more consistent expression across genes when compared to a single RBS and can be used to modify expression level to within twofold of the target range,(*10*) suggest that the T^2RMR libraries will give consistent expression levels across all genes.

Rapid construction of genome-scale libraries

In order to scale these readouts to evaluate fitness at the whole-genome scale, 4,890 targeting oligos (172 nt each) were designed to target every protein coding gene, pseudogene, and non-coding RNA in the *E. coli* MG1655 genome (Supporting Information Table S1). The targeting oligos were designed and synthesized as described in Methods. The pool of targeting oligos was amplified and ligated to each of the four different BCD modules, creating four synDNA pools. All four pools were constructed by one researcher in one week, which is approximately 2,000 times faster than using standard techniques. The libraries were then recombineered into cells in a single day, generating thousands of recombinant colonies for each of the libraries.

For this study we chose to use molecular barcodes that are optimized for high throughput sequencing to track alleles rather than using microarray. Each of the twelve-nucleotide (nt) barcodes that were used differed from every other barcode by a minimum of four nt. High throughput sequencing allows deeper coverage (up to ~10^9 for HiSeq vs ~10^4 dynamic range for microarrays(*13*)), more precise analysis of genotype frequencies, and tracking of rare alleles.(*14*) High throughput sequencing also results in lower error rates by increasing the resolution of allele counting from the typical 25-100 nt level (microarray probe lengths) down to the single nt level. This increased resolution not only improves accuracy of allele quantitation but also reveals other errors that may be present in synDNA, such as DNA synthesis errors or mismatches between the homology region and the assigned barcode sequence for a gene that arise as a result of recombination during PCR amplification of the library.(*15*)

Deep sequencing using the Illumina MiSeq instrument (http://www.illumina.com/systems.ilmn) was used to verify the coverage and uniformity of each of the T^2RMR libraries. The frequency of each barcode was calculated using a Python script by dividing the number of reads for each barcode by the total number of reads passing the initial quality filters. Over 99% of the unique gene barcodes were identified in the initial synDNA pools and in the pools from cells after multiplex recombineering (Figure 2A-B). While most barcodes are centered around the same number of counts, some barcodes were detected at unexpectedly high frequencies following recombineering and population outgrowth for unknown reasons (Figure 2C).

As a control, the synDNA pools included 100 cassettes with homology arms corresponding to the *Saccharomyces cerevisiae* genome. Since these cassettes should not be incorporated into the *E. coli* genome, they serve as a measure of noise in the sequencing data. Barcodes corresponding to yeast homology arms were generally observed at a frequency of ~10^{-4} post cellular integration (which corresponds to fewer than 5 counts for the growth selection data), indicating that yeast homology arms may be integrated into the *E. coli* genome at low levels or that errors during DNA synthesis, PCR amplification and/or sequencing result in a low rate of misidentified barcodes. To further examine the incidence of misidentified barcodes, we looked for barcode sequences in the MiSeq data that did not correspond to a designed barcode. We found that erroneous barcodes comprised 5-6% of the reads for each of the four libraries, which is similar to a previously reported error rate for high throughput oligonucleotide synthesis.(*16*) Since ~96% of these erroneous barcodes appeared with fewer than five counts (with an average of three counts), five counts per barcode was subsequently used as a threshold for excluding genes from further analysis (Figure 2D).

Validation of genome-scale fitness mapping

We next grew *E. coli* in LB rich medium or MOPS minimal medium and mapped genes for which altered expression influenced growth (Figure 3A-B). Fitness for each gene expression variant was determined using the equation $W_x = F_{x,f}/F_{x,i}$, where fitness (W) is the ratio of the final gene frequency to the initial gene frequency as determined by barcode counts. To determine whether the barcode sequences themselves affect the fitness of a gene, 100 genes were assigned multiple unique barcodes (Supporting Information Figure S1). The median coefficient of variation for these 100 genes was 0.266, with a range of 0.003-1.00. If more stringent quality filtering is applied (increasing the minimum count threshold to 50 counts), the median coefficient of variation for these 100 genes drops to 0.130, with a range of 0.005-0.629. Notably this includes all other sources of variation such as cell-to-cell fitness differences, sample preparation, and sequencing. We additionally examined whether modifying genes in the same operon resulted in similar fitness effects (Supporting Information Figure S2). When comparing genes in the same operon, the median coefficient of variation was 0.241, with a range of 0.0002-0.866, suggesting that genes in the same operon generally behave similarly when modified.

Changing the expression level of some genes does confer a fitness advantage, particularly when genes are overexpressed in minimal medium (although, as expected, the majority of genes do not confer a growth advantage in either rich or minimal media) (Figure 3A and Supporting Information Figure S3). The genes that promote growth when expressed at high levels are genes involved in cell membrane structure/composition as shown in Supporting Information Table S2. When cells are grown in rich medium, very few alleles result in a growth advantage, particularly with the highest expressing constructs. This paucity of fitness-improving genes may be due to a metabolic burden on the cells: when bacteria are required to produce large quantities of protein, growth rate tends to decrease.(*17, 18*) Alternately, these cells may already be evolved for optimal growth in rich medium and therefore few increases in fitness were observed. Despite the challenges in detecting differences in gene response to such small selective pressures as MOPS and LB, T^2RMR does significantly better than the original TRMR library for discriminating among growth medium for both library type and IPTG concentration (Figure 4).

Even though changing the expression of most genes does not confer a fitness advantage, some interesting patterns can be seen for the genes that do confer an advantage. For example when *wecF*, which catalyzes a step in outer membrane glycolipid synthesis, is expressed at any level, cells have a growth advantage in MOPS minimal medium (Figure 3A). This result agrees with a previous study showing that *wecF* expression increases when *E. coli* are grown continuously in glucose limited conditions.(*19*) On the other hand, *cdh*, which is involved in phospholipid metabolism by degrading CDP-diacylglycerols, only confers a fitness advantage in the middle expression level range but not at the highest expression level (Figure 3A). To the best of our knowledge, increasing the expression of *cdh* has not been previously reported to confer a growth advantage. Furthermore, *cdh* would likely not have been identified using a binary "on"/"off" screen, since *cdh* only confers a growth advantage over a limited expression range. Of note, our results with the "off" library confirm previously published results(*20, 21*) showing that inactivation of *cdh* does not affect cell growth.

Further characterization of T^2RMR library growth selections

For further validation of the function of the T^2RMR libraries we next evaluated the fitness of "off" library alleles essential for growth when cultured in minimal medium (Figure 3B and Supporting Information Figure S4). The majority of genes that are known to result in auxotrophy (http://cgsc.biology.yale.edu/Auxotrophs.php) were found to have a much lower fitness when cells were grown in MOPS minimal medium than when cells were grown in LB rich medium (Figure 5). These results indicate that we are able to use T^2RMR to identify genes that are known to be essential and confirms that our "off" libraries are resulting in the loss of gene function as expected. By using T^2RMR, a single researcher was able to reproduce years worth of studies on individual gene functions in under a week, highlighting the efficiency of this technique.

Finally, we demonstrated the use of the T^2RMR libraries to generate new understanding, such as through the rapid mapping of genes to traits across a broad range of conditions, as has been the focus of recent chemical genomics efforts.(*22, 23*) Specifically, we applied various multivariate analysis techniques to the T^2RMR fitness data to identify patterns among the various libraries and selections (rich or minimal medium and 4 IPTG levels). Sample fitness profiles clustered by BCD (library) type regardless of IPTG induction level and whether the libraries were grown in rich on minimal medium. This result is surprising given that our prior validation of BCD function (Fig. 1b) and a ~10,000 fold range of expression levels. These data may indicate that BCD selection tends to overwhelm the effect of IPTG induction under less stringent selection; combining stronger selective pressures with higher resolution sequencing would likely result in a more obvious expression level-fitness clustering. If the fitness landscape is sufficiently rough, however, recovering a linear expression-fitness clustering would be unlikely.

To further explore this result we applied a more stringent filter (50 reads minimum) to remove low-abundance barcodes. At this more stringent filter, clustering based on expression level became stronger, especially in minimal medium (Figure 6). In particular the induced “high” expression libraries form a distinct cluster, as do the “off” libraries. All other libraries, which all exhibited intermediate expression levels are then grouped together. This pattern is lost when the libraries are cultured in LB medium, again suggesting that the cells may already be optimized for rapid growth in rich medium and that the generally negative effects on fitness (as shown if Fig. 3b) are more likely a result of insertion of a foreign promoter (and a dual BCD) then differences in expression of the downstream genes and that such effects are what is driving the clustering results.

Conclusions

Using T^2RMR (libraries available upon request), we were able to simultaneously map the effect of changes in gene expression spanning a ~10^4 range onto fitness for every gene in the *E. coli* genome. Notably, a single researcher was able to screen these ~20,000 gene variants in under a week. Through this approach, we identified genes that confer a fitness advantage or disadvantage at different ranges of expression levels in two different growth environments. Furthermore, we have shown that the “off” library functions as a barcoded deletion library, which can be used to screen for genes that are essential in any growth condition (e.g. auxotrophy).

Precisely-tuned promoter and BCD insertion combined with highly quantitative fitness tracking allows detection of new genotypes and phenotypes that could not be identified by other methods, allowing researchers to rapidly explore more types of genome-scale genotype-phenotype connections more rapidly than would otherwise be possible. T^2RMR therefore represents a substantial advance for the rapid engineering of strains with specified target phenotypes by providing the standardized parts that are required for systems engineering, especially for phenotypes requiring coordinated action across multiple genes or pathways that are sensitive to small changes in expression levels. Future work will examine applications of T^2RMR to tolerance and production phenotypes, such as antibiotic resistance and genome-scale screening for loci affecting production of biofuels, terpenoids, and other metabolites of interest.

METHODS

Strains and DNA

All experiments were performed in wild-type *Escherichia coli* MG1655 (ATCC 700926). Genomic sequences for the oligonucleotide library were obtained from Genbank record U00096.2 and were annotated using the Ecogene database version 3.0 (http://www.ecogene.org/)(*24, 25*). Protein coding genes, pseudogenes, and non-coding RNAs were all included in the library. The oligonucleotide library was purchased from Agilent. The BCD sequences and surrounding elements were synthesized by Genewiz. The pKD13 plasmid(*26*) was acquired from the Coli Genetic Stock Center (CGSC) (7633) and was used as a template for PCR amplification of the kanamycin resistance cassette. The pSIM5 plasmid(*27*) was a gift from D. Court.

β-galactosidase assays

β-galactosidase assays were performed as previously described(*28*) with minor modifications. Briefly, cells were inoculated into 5 ml LB medium with different concentrations of IPTG (0.01 mM, 0.125 mM, or 1 mM) or no IPTG. Cells were grown at 37 °C until they reached an optical density at 600 nm of 0.5-0.7. Cells (1 ml) were then collected by centrifugation and the supernatant was removed. The pellets were resuspended in 100 μl PopCulture Reagent (EMD Millipore) and 1 μl lysozyme to lyse the cells. In a 96-well plate, 1-10 μl cell lysate was combined with 132 μl Z-buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0**)**. Ortho-Nitrophenyl-β-galactoside (ONPG; 15 or 29 μl) was added to each well and the reactions were incubated at room temperature until a faint yellow color was observed, at which point 75 μl Na_2CO_3 was added to quench the reaction. Yellow color was quantitated by measuring the A_{420} value for each sample. All assays were performed in triplicate. β-galactosidase activity was calculated using the following equation:

$$\text{Modified Miller units} = 100 \times \frac{\text{Abs}_{420}}{(\text{Abs}_{600})(\text{time})\left(\frac{\text{volume lysate}}{\text{total volume}}\right)}$$

Library construction and recombineering

Library construction is shown in Supporting Information Figure S5**.** Targeting oligos (4,890 unique 172-mers) were designed to target every protein coding gene, pseudogene, and non-coding RNA in the *E. coli* MG1655 genome (Supporting Information Table S1). Each oligo contains: flanking regions for PCR amplification (23 nt for 5', 31 nt for 3'), downstream homology region (40 nt), AscI site (8 nt), upstream homology region (40 nt), barcode priming site (18 nt), barcode sequence (12 nt). The barcodes were designed so that each barcode varies from every other barcode by at least 4 nt, has between 40 and 60% GC content, and the final barcodes are sorted to maximize heterogeneous nucleotide signals. The Python script for barcode generation and sorting is available at this URL: https://gist.github.com/walterst/f20f41e1c8385ef27620. The homology arms were designed to insert the cassettes immediately upstream of each gene and replace the native start codon. The cassettes also include a three frame stop codon to prevent read-through from any upstream gene. In a small number of cases, insertion of the cassette may result in truncation of an upstream gene product.

Shared DNAs were designed to regulate the expression level of each gene. Each shared DNA contains a kanamycin resistance gene flanked by FRT sites, three frame stop codons, a terminator spacer, a terminator pause, a promoter insulator, a LacI regulated synthetic inducible promoter (apFAB906), bicistronic design (dual) RBS, and a barcode identifying the BCD. Four different BCD sequences were chosen for each of the libraries to span a range of expression levels(*10*). The BCD sequences used are given below with annotations as follows: Red bold = putative Shine-Dalgarno motifs; grey text highlighter = short peptide coding 1st cistron; yellow text highlighter = overlapping stop codon (for the short peptide) and start codon (for the downstream gene); blue TAA = premature termination of short peptide:

"Off" (BCD22-taa),
GGGCCCAAGTTCACTTAA**AAAGGAGAT**CAACAATGAAAGCAATTTTCtaaCTGAAACATCTTAATCATGC**CT**
AGGAAGTTTTCTAATG
"Low" (BCD21),
GGGCCCAAGTTCACTTAA**AAAGGAGAT**CAACAATGAAAGCAATTTTCGTACTGAAACATCTTAATCATGC**G**
AGGGATGGTTTCTAATG
"Intermediate" (BCD19),
GGGCCCAAGTTCACTTAA**AAAGGAGAT**CAACAATGAAAGCAATTTTCGTACTGAAACATCTTAATCATGC**T**
ATGGAGGTTTTCTAATG
"High" (BCD2),
GGGCCCAAGTTCACTTAA**AAAGGAGAT**CAACAATGAAAGCAATTTTCGTACTGAAACATCTTAATCATGC**T**
AAGGAGGTTTTCTAATG

The BCD sequences and surrounding elements were ligated to the kanamycin resistance cassette by the Gibson assembly method (NEB) to create the shared DNA cassettes. A portion of targeting oligonucleotides were amplified by PCR and then ligated with the shared DNA cassettes using the Golden Gate cloning method(*29*). Ligation products were then amplified by rolling circle amplification(*30*), treated with AscI and mung bean nuclease, and purified to generate the synDNA cassettes. All reactions were carried out in parallel with each of the four shared DNA cassettes.

Recombineering was performed as previously described(*31*), with minor modifications. Briefly, *E. coli* MG1655 cells containing the pSIM5 plasmid were grown in 400 ml low-salt LB at 30 °C. When cells reached an optical density at 600 nm of 0.6-0.7, the culture flask was transferred to a 42 °C shaking water bath for 15 minutes to induce the λ–Red proteins. The flask was then placed immediately on ice and all subsequent steps were carried out at 4 °C. The cells were collected by centrifugation and then washed twice in ice-cold deionized water. The cells were resuspended in a final volume of 1.6 ml water. Cells were divided into 50 µl aliquots and were transformed with approximately 300 ng of synDNA by electroporation with a pulse of 18 kV cm^{-1}. Electroporation was carried out eight times for each of the four synDNA libraries. The transformed cells from each library were combined in 50 ml SOC medium and were allowed to recover for 6 h at 30 °C. Cells were then collected by centrifugation, resuspended in 4 ml LB, spread on LB agar plates containing kanamycin (100 µg ml^{-1}). Plates were incubated overnight at 37 °C. Colonies were scraped from the agar plates and resuspended to $6x10^9$ cells per milliliter in LB medium containing 15% (vol/vol) glycerol. Aliquots of each library were stored at -70 °C.

Growth selections

Freezer stocks were used to inoculate 10 ml LB medium with each of the four libraries. Cells were grown at 37 °C for approximately 45 min. These cultures were then diluted to 2 x 10^8 cells per milliliter in LB medium containing kanamycin (100 µg ml^{-1}) and all four libraries were mixed together in equal amounts. The mixed libraries were then aliquoted into 10 ml volumes in LB medium containing kanamycin (100 µg ml^{-1}) and different concentrations of IPTG (0.01 mM, 0.125 mM, or 1 mM) or no IPTG. These cultures were grown at 37 °C for approximately 40 min to an optical density at 600 nm of 0.6. A 1 ml aliquot of each culture was centrifuged and the pellet was frozen for sequencing analysis. The remainder of the cells were then collected by centrifugation, washed once in 1x MA salts (5x: 1 liter H_2O, 52.5 g K_2HPO_4, 22.5 g KH_2PO_4, 5 g $(NH_4)_2SO_4$, 2.5 g sodium citrate·$2H_2O$), and resuspended in 10 ml 1x MA salts. The washed cells (3 x 10^5 cells) were used to inoculate 10 ml of MOPS minimal medium with 4% (wt/vol) glucose or LB medium, each containing a different concentration of IPTG or no IPTG. Cells were grown at 37 °C until they reached an optical density at 600 nm of 1.0-1.3 (approximately 8 generations). A 1 ml aliquot of each culture was centrifuged and the pellet was frozen for sequencing analysis.

Sequencing and Fitness calculations

Frozen cell pellets were resuspended in deionized water and lysed by boiling. The clarified lysate (2-8 µl), which contains genomic DNA, was used as a template for PCR to amplify the molecular barcodes in the synDNA cassettes. A second round of PCR was performed to add Illumina adapter sequences flanking the synDNA amplicon. High throughput sequencing was performed on an Illumina MiSeq using 300-cycle MiSeq Reagent Kits v2 (http://www.illumina.com).

USEARCH(*32*) was used for quality filtering of the raw MiSeq fastq files. All sequence reads shorter than 80 nucleotides or with a quality score below 15 were excluded from further analysis. Barcode frequencies were calculated using a Python script by dividing the number of reads for each individual barcode, normalized by the total number of reads in the initial and final libraries. Barcodes that did not meet a minimum count threshold (either 5 or 50 reads) were also excluded from analysis. Fitness values were calculated by dividing the initial barcode frequency by the final barcode frequency. Negative fitness values were calculated by taking the $\log_2$ of the positive fitness values.

Genome plots were generated using the Circos software package(*33*). Genes were plotted by genomic locus on the circular *E. coli* chromosome. Depending on the plot, the height of the spike at each location represents either the frequency of that gene in the MiSeq data or the fitness of that gene.

Data analysis

For analysis of the behavior of genes in the same operon, the curated operon organization was obtained from RegulonDB (http://regulondb.ccg.unam.mx/). The average of the fitnesses for each gene, the standard deviations, and coefficients of variance were calculated using Python. The list of gene deletions leading to auxotrophy was obtained from the Coli Genetic Stock Center (http://cgsc.biology.yale.edu/Auxotrophs.php).

Principal coordinates analysis was performed using QIIME(*34*) and EMPeror(*35*) software packages. Non-parametic ANOVA (PERMANOVA) was done using the 'adonis' function in the 'vegan' R package(*36*). Hierarchical clustering was performed using R(*37*) with Pearson correlation for the distance metric and complete linkage clustering. The input data was the fitness for each gene in each library for each of the four IPTG concentrations. Fitness data in LB rich medium and MOPS minimal medium were analyzed separately.

Microarray data from the original TRMR libraries were processed exactly according to Warner et al(*2*).

SUPPORTING INFORMATION

Tables S1-S2, Figures S1-S5.

ACKNOWLEDGEMENTS

We thank the Gill lab and J. Warner for helpful discussions. We thank W.A. Walters for the barcode generation software. We also thank D. Court (NIH) for providing the pSIM5 plasmid. This research was supported by the Office of Science (BER), U. S. Department of Energy, DE-SCOOO8812.

1. Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., and Church, G. M. (2009) Programming cells by multiplex genome engineering and accelerated evolution, *Nature 460*, 894-898.
2. Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B., and Gill, R. T. (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides, *Nature biotechnology 28*, 856-862.
3. Wang, X., Yomano, L. P., Lee, J. Y., York, S. W., Zheng, H., Mullinnix, M. T., Shanmugam, K. T., and Ingram, L. O. (2013) Engineering furfural tolerance in Escherichia coli improves the fermentation of lignocellulosic sugars into renewable chemicals, *Proceedings of the National Academy of Sciences of the United States of America 110*, 4021-4026.
4. Raman, S., Rogers, J. K., Taylor, N. D., and Church, G. M. (2014) Evolution-guided optimization of biosynthetic pathways, *Proceedings of the National Academy of Sciences of the United States of America 111*, 17803-17808.
5. Zeitoun, R. I., Garst, A. D., Degen, G. D., Pines, G., Mansell, T. J., Glebes, T. Y., Boyle, N. R., and Gill, R. T. (2015) Multiplexed tracking of combinatorial genomic mutations in engineered cell populations, *Nature biotechnology*.
6. Winkler, J. D., and Kao, K. C. (2014) Recent advances in the evolutionary engineering of industrial biocatalysts, *Genomics 104*, 406-411.
7. Glebes, T. Y., Sandoval, N. R., Gillis, J. H., and Gill, R. T. (2015) Comparison of genome-wide selection strategies to identify furfural tolerance genes in Escherichia coli, *Biotechnology and bioengineering 112*, 129-140.
8. Sandoval, N. R., Kim, J. Y., Glebes, T. Y., Reeder, P. J., Aucoin, H. R., Warner, J. R., and Gill, R. T. (2012) Strategy for directing combinatorial genome engineering in Escherichia coli, *Proceedings of the National Academy of Sciences of the United States of America 109*, 10540-10545.
9. Temme, K., Zhao, D., and Voigt, C. A. (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca, *Proceedings of the National Academy of Sciences of the United States of America 109*, 7085-7090.

10. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements, *Nature methods 10*, 354-360.
11. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics, *Analytical chemistry 76*, 4193-4201.
12. Gilchrist, A., Au, C. E., Hiding, J., Bell, A. W., Fernandez-Rodriguez, J., Lesimple, S., Nagaya, H., Roy, L., Gosline, S. J., Hallett, M., Paiement, J., Kearney, R. E., Nilsson, T., and Bergeron, J. J. (2006) Quantitative proteomics analysis of the secretory pathway, *Cell 127*, 1265-1281.
13. Pierce, S. E., Fung, E. L., Jaramillo, D. F., Chu, A. M., Davis, R. W., Nislow, C., and Giaever, G. (2006) A unique and universal molecular barcode array, *Nature methods 3*, 601-603.
14. Robins, W. P., Faruque, S. M., and Mekalanos, J. J. (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene, *Proceedings of the National Academy of Sciences of the United States of America 110*, E848-857.
15. Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990) DNA recombination during PCR, *Nucleic acids research 18*, 1687-1691.
16. Kosuri, S., Eroshenko, N., Leproust, E. M., Super, M., Way, J., Li, J. B., and Church, G. M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips, *Nature biotechnology 28*, 1295-1299.
17. Bentley, W. E., Mirjalili, N., Andersen, D. C., Davis, R. H., and Kompala, D. S. (1990) Plasmid-encoded protein: the principal factor in the "metabolic burden" associated with recombinant bacteria, *Biotechnology and bioengineering 35*, 668-681.
18. Hoffmann, F., and Rinas, U. (2001) On-line estimation of the metabolic burden resulting from the synthesis of plasmid-encoded and heat-shock proteins by monitoring respiratory energy generation, *Biotechnology and bioengineering 76*, 333-340.
19. Franchini, A. G., and Egli, T. (2006) Global gene expression in Escherichia coli K-12 during short-term and long-term adaptation to glucose-limited continuous culture conditions, *Microbiology 152*, 2111-2127.
20. Bulawa, C. E., and Raetz, C. R. (1984) Isolation and characterization of Escherichia coli strains defective in CDP-diglyceride hydrolase, *The Journal of biological chemistry 259*, 11257-11264.
21. Babinski, K. J., Ribeiro, A. A., and Raetz, C. R. (2002) The Escherichia coli gene encoding the UDP-2,3-diacylglucosamine pyrophosphatase of lipid A biosynthesis, *The Journal of biological chemistry 277*, 25937-25946.
22. Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C., and Giaever, G. (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes, *Science 320*, 362-365.
23. Pathania, R., Zlitni, S., Barker, C., Das, R., Gerritsma, D. A., Lebert, J., Awuah, E., Melacini, G., Capretta, F. A., and Brown, E. D. (2009) Chemical genomics in Escherichia coli identifies an inhibitor of bacterial lipoprotein targeting, *Nature chemical biology 5*, 849-856.
24. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12, *Science 277*, 1453-1462.
25. Rudd, K. E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12, *Nucleic acids research 28*, 60-64.
26. Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products, *Proceedings of the National Academy of Sciences of the United States of America 97*, 6640-6645.
27. Datta, S., Costantino, N., and Court, D. L. (2006) A set of recombineering plasmids for gram-negative bacteria, *Gene 379*, 109-115.
28. Thibodeau, S. A., Fang, R., and Joung, J. K. (2004) High-throughput beta-galactosidase assay for bacterial cell-based reporter systems, *BioTechniques 36*, 410-415.

29. Engler, C., Kandzia, R., and Marillonnet, S. (2008) A one pot, one step, precision cloning method with high throughput capability, *PloS one 3*, e3647.
30. Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification, *Genome research 11*, 1095-1099.
31. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G., and Court, D. L. (2009) Recombineering: a homologous recombination-based method of genetic engineering, *Nature protocols 4*, 206-223.
32. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics 26*, 2460-2461.
33. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009) Circos: an information aesthetic for comparative genomics, *Genome research 19*, 1639-1645.
34. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data, *Nature methods 7*, 335-336.
35. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013) EMPeror: a tool for visualizing high-throughput microbial community data, *GigaScience 2*, 16.
36. Oksanen, J., Guillaume Blanchet, F., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2015) Community Ecology Package: R package version 2.2-1.
37. The R Core Team. (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

For Table of Contents Use Only

Genome-wide tuning of protein expression levels to rapidly engineer microbial traits

Emily F Freed, James D Winkler, Sophie J Weiss, Andrew D Garst, Vivek K Mutalik, Adam P Arkin, Rob Knight, Ryan T Gill

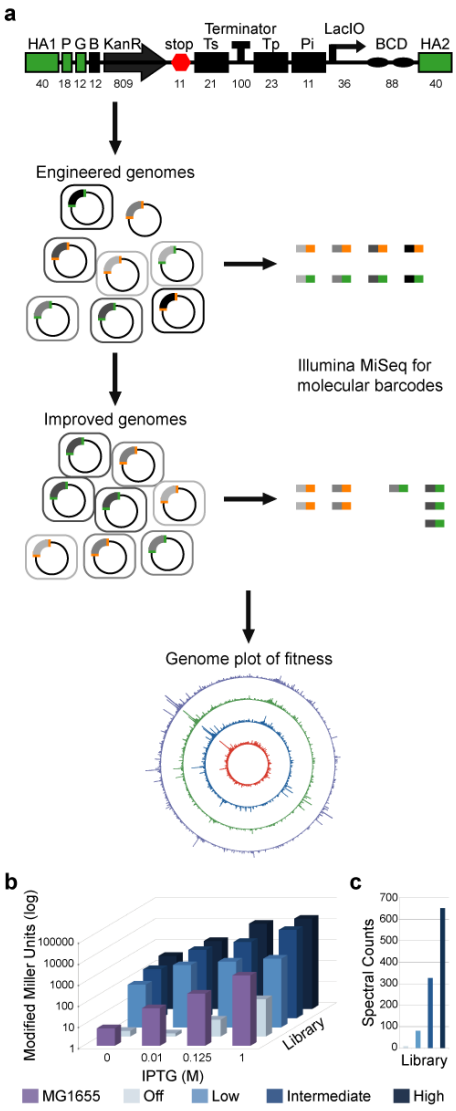

Figure 1: T^{2}RMR strategy for genome engineering. (A) T^{2}RMR cassettes consist of multiple modules assembled from targeting oligos (green) and shared DNA (black). Each cassette specifically modifies expression of a single gene. HA1 and HA2, homology regions; P, barcode priming site; G, barcode identifying the gene; B, barcode identifying the BCD; KanR, kanamycin resistance gene; stop, three frame stop codons; Ts, terminator spacer; Tp, terminator pause; Pi, promoter insulator; LacIO, LacI regulated synthetic inducible promoter (apFAB906); BCD, bicistronic design (dual RBS). Numbers underneath indicate the size of each part in base pairs. Cassettes are chromosomally integrated into *E. coli* cells via recombineering, creating a pool of cells that each has a specific gene modified. Recombineered cells then are screened or selected for the desired phenotype. High throughput sequencing of barcodes is used to calculate the fitness of each allele and to map specific changes to the trait of interest. (B) Miller assays show that *lacZ* alleles from each of the four libraries give the intended phenotype and cover a ~10^4 range in expression level. The inset shows values below 400 modified Miller units. (C) Mass spectrometry shows that the β-galactosidase protein is expressed over a ~10^2 range and each of the four libraries behaves as expected.

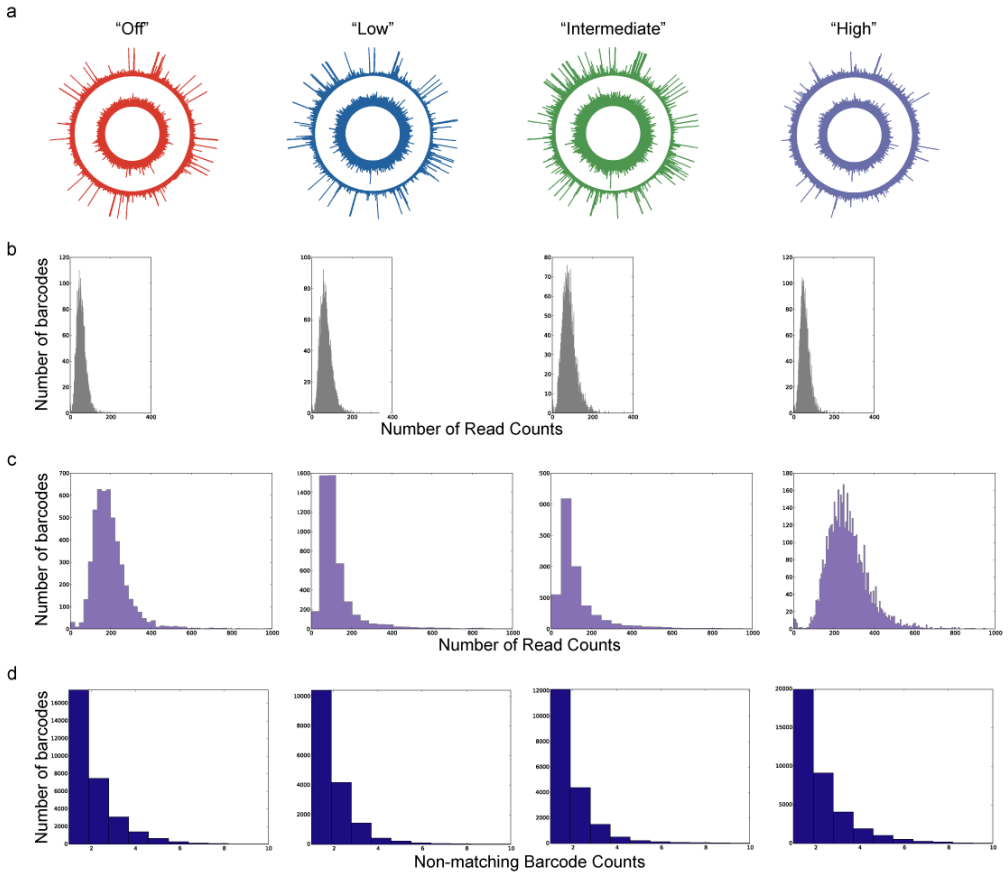

Figure 2. Analysis of synDNA and cell libraries. (A) Distribution of synDNA (inner circles) or alleles after incorporation into *E. coli* cells (outer circles) plotted by genomic location on the circular chromosome. Red, "Off" library; blue, "Low" library; green, "Intermediate" library; purple, "High" library. (B) Histograms showing the distribution of barcodes for each synDNA library as determined by high throughput sequencing. (C) Histograms showing the distribution of barcodes for each cellular library as determined by high throughput sequencing. (D) Histograms showing the distribution of erroneous barcodes for each cellular library as determined by high throughput sequencing.

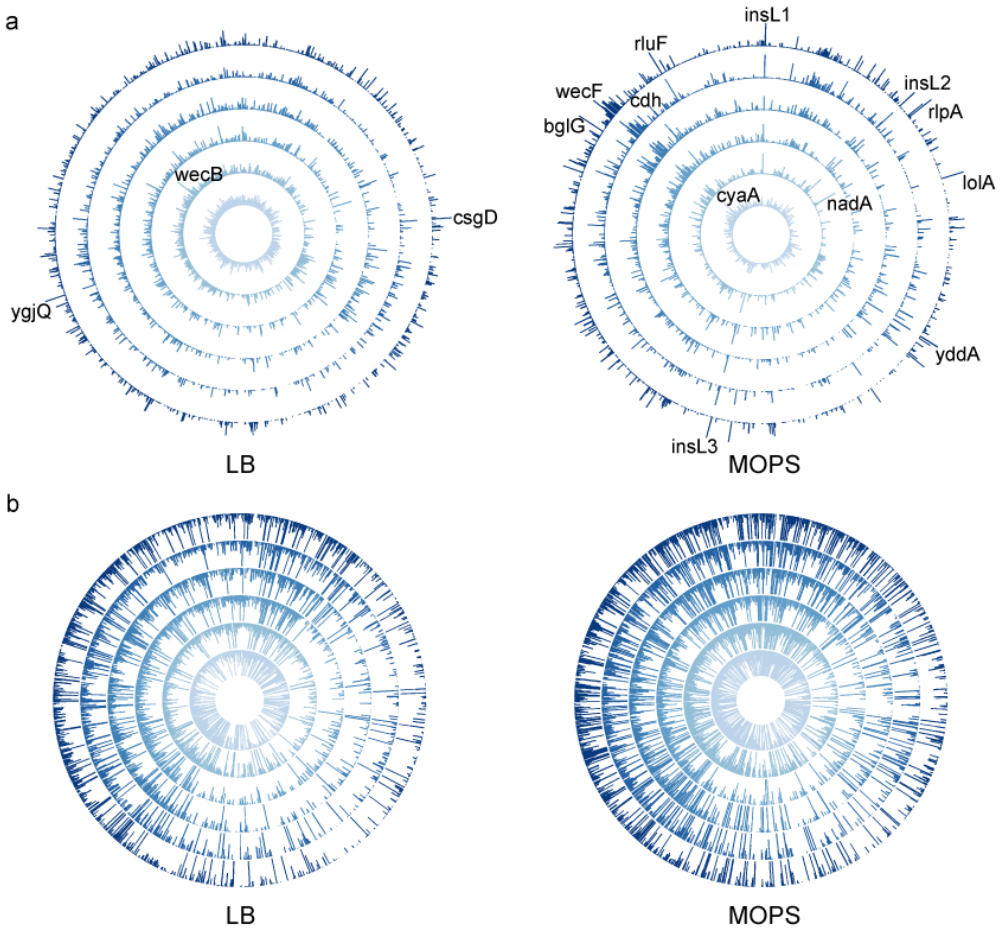

Figure 3. Identification of alleles after growth in LB or MOPS. Fitness for each allele was calculated by barcode frequency detected by high throughput sequencing. The $\log_2$ value of fitness was plotted by genomic location. Alleles with positive fitness values are shown in (A) and alleles with negative fitness values are shown in (B). Conditions from the inner to the outer circle: “Off” library with 0 mM IPTG, “Off” library with 0.125 mM IPTG, “Low” library with 0.125 mM IPTG, “Intermediate” library with 0.125 mM IPTG, “High” library with 0.125 mM IPTG, “High” library with 1 mM IPTG.

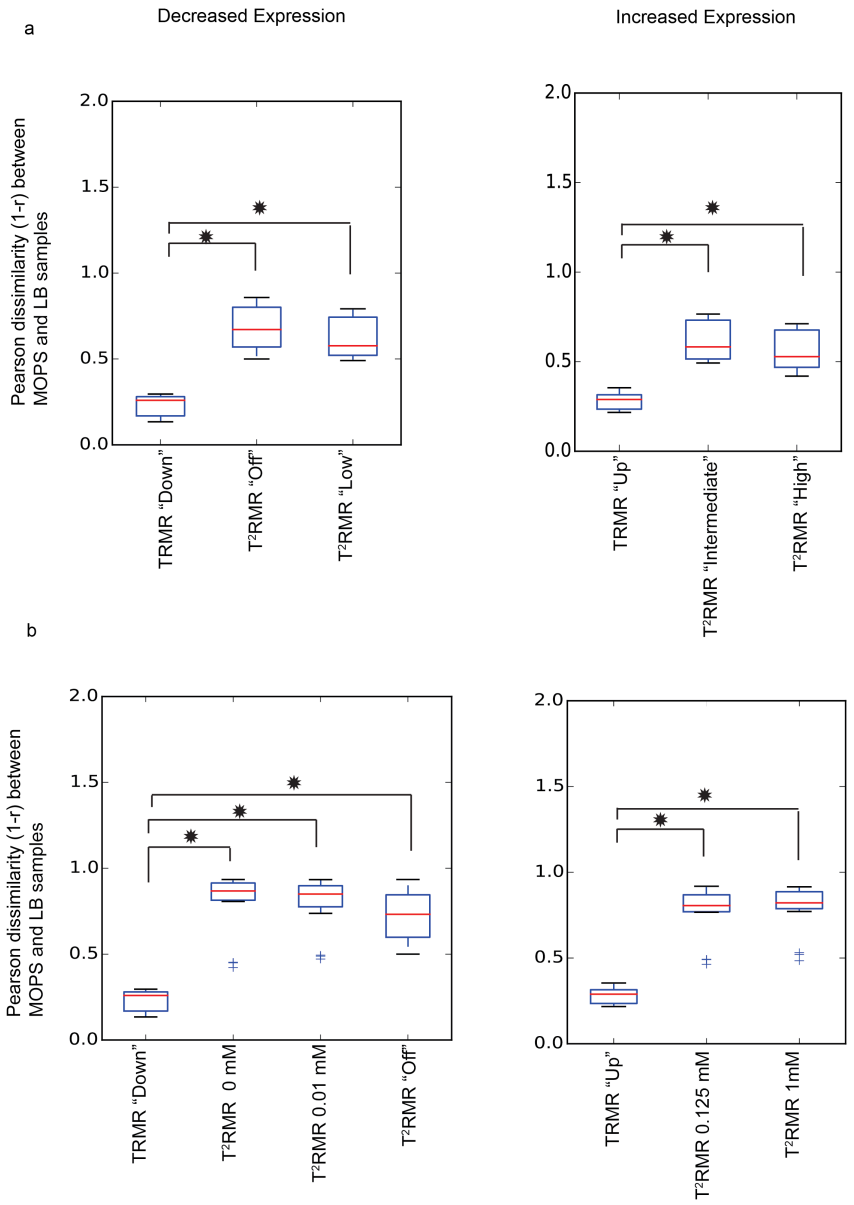

Figure 4. T^{2}RMR has significantly increased ability to discriminate between selective pressures, such as MOPS minimal and LB rich medium. The Pearson dissimilarity (0 indicates perfectly linearly correlated, and 2 indicates negatively correlated) between MOPS and LB samples for (A) each library type and (B) each IPTG concentration. * indicates $p < 0.05$ Benjamini-Hochberg corrected significance. All IPTG concentrations for the T^{2}RMR "Off" library were grouped together, since the "Off" library is always off.

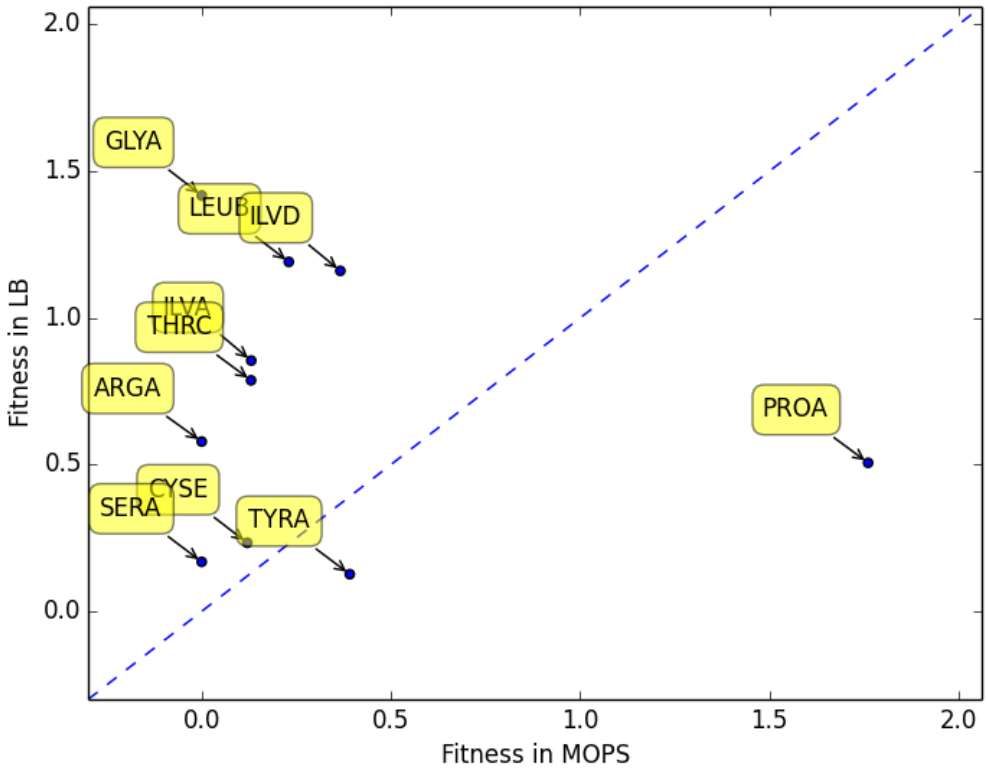

Figure 5. Analysis of auxotrophy. The fitness of genes that are known to cause auxotrophy when deleted was compared in the "off" libraries for cells grown in LB rich medium vs. cells grown in MOPS minimal medium. Any value less than 1.0 indicates turning off that gene using T^2RMR results in a fitness disadvantage in that medium.

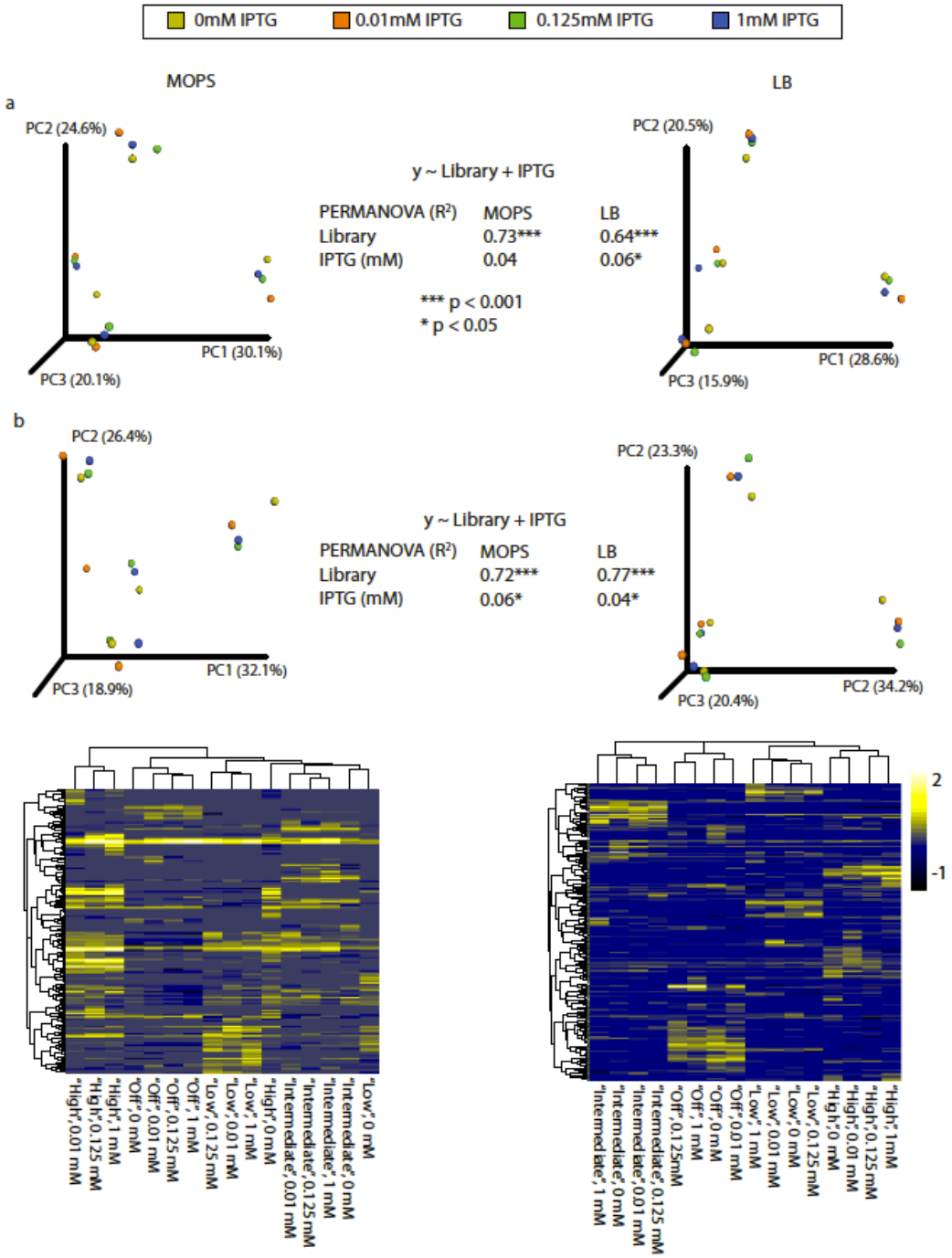

Figure 6. Multivariate analysis. (A) Principal Coordinates Analysis (PCoA), and associated non-parametric multivariate ANOVA (PERMANOVA), of fitness values after growth in MOPS minimal medium (left) or LB rich medium (right) using 5 initial counts as a threshold. (B) PCoA, the associate PERMANOVA, and hierarchical clustering of fitness values after growth in MOPs minimal medium (left) or LB rich medium (right) using 50 initial counts as a threshold. All clustering was done with Pearson correlation, and genes whose values were all zeroes were removed.